Research Summary – Final Report
Karsten Schwan, PI, Greg Eisenhauer, Matt Wolf, Co-PIs

Data Staging on Future Platforms: Systems Management for High Performance and Resilience

Motivation and Background. Well-known I/O bottlenecks for high end machines have engendered research and development in which the outputs generated by scientific simulations are moved to 'staging areas' or 'burst buffers', where they are organized, reduced, and analyzed before they are moved to backend parallel file systems. The resulting I/O pipelines, however, must be 'managed' in order to maintain the processing and output rates required to avoid applications from blocking on output. Specific management tasks include (1) output/communication scheduling to prevent output actions from interfering, on the shared interconnect, with application-level communications, (2) right-sizing analysis to cope with varying processing needs and output data changes (e.g., interesting features exist in said data) and/or as output volumes change due to increased computational resolution (e.g., in AMR codes), and (3) coping with failures (e.g., software failures) in the analysis or visualization codes applied to dynamic simulation outputs.

Work Completed. We have developed an initial implementation of a system abstraction in which flexible management actions can be associated with componentized I/O pipelines. Each component, such as a MPI code running some parallel data analysis, is enclosed in a management abstraction termed a 'container', the role of which is to provide the resources needed to run the component. Jointly, the set of containers supporting components are driven by performance constraints concerning the entire pipeline, such as its end to end latency. The 'containerized' runtime was implemented for high end machines based on the DataTap communication middleware, and it was evaluated with a representative I/O pipeline for the LAMMPS code. Performance evaluations show that with containers, high throughput pipelines can be obtained with little additional overhead due to the use of this abstraction and with advantages derived from the fact that end to end latency can be maintained even when analytics actions in containers change.

In separate work, we have collaborated with Sandia to develop efficient transaction protocols to provide guarantees for the data being moved when I/O is performed.

Outcomes. Our Sandia collaborators with our PhD student involved published a paper on the transactional support. Karsten Schwan presented the concept at the by invitation only DOE ExaOSR workshop in Washington, DC, in Sept. 2012. A paper was submitted to the IPDPS conference (pending). A major reimplementation of the container software was completed to permit it to be run on next generation petascale machines like Titan at ORNL.
ompute state data.